

ソーシャルタギングシステムを用いたユーザの関心の発見

レオン末松豊インティ[†] 村里 英樹[†] 藤原 義久^{††,†††} 木俣 豊[†]

[†] 情報通信研究機構

^{††} ATR メディア情報科学研究所

^{†††} 京都大学総合人間学部

E-mail: [†]yutaka.leon@acm.org, ^{††}{murasakihitode,yoshi.fujiwara}@gmail.com, ^{†††}kidawara@nict.go.jp

あらまし 多数の人間と情報が共有できるソーシャルタギングシステムでは、ユーザが関心を持つコンテンツの内容を反映したタグが付与される。本論文では、ユーザが共有する関心を発見するためのタグ間の階層的なクラスタリング手法を提案する。評価実験として delicious のデータを利用し、提案手法の有効性を確認した。

キーワード ソーシャルタギングシステム, 階層的なクラスタリング, 再帰的コミュニティ抽出

Exploring User Interest in Social Tagging Systems

Yutaka I. LEON-SUEMATSU[†], Hideki MURASATO[†], Yoshi FUJIWARA^{††,†††}, and Yutaka KIDAWARA[†]

[†] National Institute of Information and Communications Technology (NICT), Japan

^{††} ATR Laboratories, Japan

^{†††} Faculty of Integrated Human Studies, Kyoto University, Japan

E-mail: [†]yutaka.leon@acm.org, ^{††}{murasakihitode,yoshi.fujiwara}@gmail.com, ^{†††}kidawara@nict.go.jp

Abstract In Social Tagging Systems, where many people can share information, users tend to use descriptive tags to annotate the contents that they are interested. In this paper we present a hierarchical clustering algorithm between tags for exploration of users' shared interest. We employ data from delicious to evaluate the effectiveness of our algorithm.

Key words Social Tagging System, Hierarchical Clustering, Recursive Community Extraction

1. はじめに

インターネット上で我々は日々、様々なソーシャルサービスを通じて交流し合っている。その中で、情報共有システムとして、ソーシャルタギングシステムが注目されている。ソーシャルタギングシステムは、ユーザが興味や関心を持つコンテンツを登録し、自由に複数のタグを付与できるシステムである。これらのコンテンツは登録したユーザにとって、何らかの価値のある情報を持つと考えられ、この情報を用いた情報検索 [2] [7], 情報推薦 [13] などの研究が盛んである。

本論文では、ソーシャルタギングシステムの一つである、ソーシャルブックマークサービスを扱う。ソーシャルブックマークサービスは、オンライン上でウェブページのブックマークを不特定多数の人間と共有することができるサービスである。ユーザは URL に対してタグを付けて分類する。各 URL に対して自由に、複数のタグを付けることができる。自分なりのタグを付けて整理したり、他人のタグやブックマークを観たり、同じ

タグで分類している他人のブックマークを確認することなどができ、同じ傾向のウェブページが見付けやすくなる。現在は、500 万以上のユーザを持つ「delicious」^(注1) が世界最大のソーシャルブックマークサービスであり、国内では「はてな」^(注2) や「Buzzurl」^(注3) がよく知られている。

これらのサービスでは入手できるデータ量が莫大なため、スパム対策、検索、推薦などの様々な課題が残されている。その中の一つの重要な課題はユーザが共有する関心の発見である。

一般に、ユーザは自分が興味を持ったページに対してタグを付ける。Li らは、ユーザが付与するタグはページの内容を表していることが多いことを報告した [9]。一方で一つのページには様々なタグが付けられる。このため、タグはユーザがページのどこに着目したかを反映していると考えられる。我々は、これ

(注1) : <http://delicious.com>

(注2) : <http://www.hatena.ne.jp>

(注3) : <http://buzzurl.jp>

を利用して、ユーザ全体の関心の傾向の発見に取り組んだ。

ユーザの関心を表すタグには詳細さの異なる様々な段階がある。従って、ユーザの利用するタグには階層構造があると考えられる。

我々は、本論文で、ユーザの関心を表すタグ間とそれらの階層的な関係性を発見するために、高速なコミュニティ抽出アルゴリズムと階層的なクラスタリング手法を提案する。

実験のため、delicious のタグ及びブックマークを対象とし、3ヶ月かけてクロールリングを行い、元データとした。

本論文の構成は次の通りである。まず、2節において関連研究を述べ、3節において提案する階層的なクラスタリング方法である再帰的なコミュニティ抽出法を説明する。続いて、4節では delicious のデータによる実験結果を示し、5節ではまとめとする。

2. 関連研究

ソーシャルブックマークサービスにおけるユーザの関心の発見に関しては、大きく三つのアプローチに分類される。

一つ目は、ユーザ中心アプローチである、Ali-Hasan ら [1] と湯田 [14] は、ユーザが張るリンク（友人、ファン）を利用して、興味が近いユーザの発見を行っている。しかし、ソーシャルネットワークにおけるリンクはユーザの興味が一致していることを表すとは限らないため、興味の異なる複数の集団が一つのグループにまとめられてしまうことがある。また、ソーシャルブックマークサービスではソーシャルネットワークサイトと異なり、ユーザ同士はリンクを張らないことが多いため、ユーザを中心した関心の解析は限界がある。

二つ目は、リソース（コンテンツ、URL）中心アプローチである。Gou ら [6] は、P2P で同じリソースを持つユーザの興味の傾向が近いと仮定し、これに基づいて、別のリソースを検索するアルゴリズムを提案した。この方法では、あるリソースのどこにユーザが興味を示したかは明示されない。また、ユーザ全体の関心の傾向を調べる目的では利用できない。

三つ目は、タグ中心アプローチである。これは、ユーザが登録したタグを利用し、ユーザの関心を発見するアプローチである。我々はこのアプローチに着目する。タグの共起からグラフを作成し、これを解決することにより、ユーザ全体の関心の階層構造とそれらの関係を発見することを目指す。

尚、扱うデータ量が大規模なため、高速なクラスタリングアルゴリズムが必要になった。我々 [8] は、コミュニティ抽出アルゴリズムの改良を加えて、従来では不可能であった数千万ノード以上の大規模な重みなしのグラフのコミュニティ抽出を可能にした。

具体的には、大規模なグラフから高速にコミュニティ抽出をするために、従来法である Newman 法 [11] の改良である Clauset らによる CNM 法 [3] の改良を行った。実装レベルの改良により、同じ CNM 法と比較して7倍の性能向上が確認できた。また、先行研究の Danon らによる DDA 法 [4] では、リンク含有量による差を均質化させることにより、分割の評価指標である Q 値の向上は報告されていたものの、速度の向上は無いとさ

れていた。DDA 法の高高速化のための改良を行った結果、従来 CNM 法では 43 時間を要した 100 万ノード規模のコミュニティ抽出を、Q 値を向上させつつ、わずか 14 分に短縮することができた。この結果、実行速度を 192 倍高速にすることができた。

今回は、タグ間の共起頻度を使用するため、[8] で提案したアルゴリズムを重み付きグラフ用に適応した。3.1 では、その抽出法について述べる。

3. 階層的なクラスタリング手法

3.1 コミュニティ抽出

グラフのクラスタリング手法の中で、大規模なグラフにも実用的に適用できるアルゴリズムの一つは、Newman [11] により導入されたグラフ全体の分割を評価する指標、モジュラリティ Q に基づく Hill-climbing アルゴリズムがある。モジュラリティ Q [12] はグラフをいくつかのコミュニティに分割したとき、その各コミュニティ内外でのリンク数の偏りをもとに算出される。全分割がどれだけグラフ全体をリンク高密度なサブグラフに分割したかを評価する指標である。重み付きのグラフの場合は以下の式にしたがって算出する。

$$Q = \sum_i (e_{ii} - a_i^2) \quad (1)$$

ここで、 e_{ii} はグラフ全体の総重みに対する、コミュニティ i 内のリンクの総重みの割合である。 a_i はグラフ全体の総重みに対する、コミュニティ内の全ノードの出・入リンクの総重みの割合である。それぞれは、次の式で得られる。

$$e_{ij} = \frac{1}{2m} \sum_u \sum_v A_{uv} \delta(c_u, i) \delta(c_v, j) \quad (2)$$

$$a_i = \sum_j e_{ij} \quad (3)$$

A_{uv} はノード u と v のリンクがある場合はそのリンクの重み、なければ 0 を返す。 m は全リンクの総重みを表す。

開始時点では、各ノードを内包ノード数が 1 である仮コミュニティとみなして、まず $Q_{initial}$ を求める。次に、すべてのリンクに対してそれぞれを結合した場合の Q の増加分 ΔQ_{ij} を求め、最大 ΔQ を与えるリンクで結ばれた仮コミュニティ同士を結合する。このプロセスを繰り返す。全ての ΔQ_{ij} が負となった時点で Q は最大となるので、ここで計算を終了する。

今回は高速化とより高い Q 値を得るために ΔQ_{ij} を以下の式での正規化する。

$$\Delta Q_{ij}^* = \frac{\Delta Q_{ij}}{\min(a_i, a_j)} \quad (4)$$

Newman 法では最大の ΔQ を得られるリンクで結ばれた仮コミュニティ同士の結合を行うが、我々は、最大の ΔQ_{ij}^* を与えるリンクで結ばれた仮コミュニティ同士の結合を行う。 ΔQ_{ij}^* は結合される仮コミュニティの選択にのみ利用し、 ΔQ の再計算は通常通りに行う。

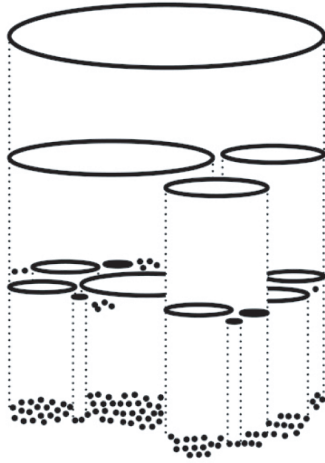


図1 再帰的コミュニティ分割
Fig.1 Recursive Community Extraction

3.2 再帰的なコミュニティ抽出

モジュラリティ Q が最大であっても、コミュニティが求めたい粒度であるとは限らない。Fortunato ら [5] は、モジュラリティ自体が原理的にもつコミュニティの解像度があり、相対的に小さいコミュニティが抽出されないと報告した。例えば、数珠のような一本のリンクで結ばれた複数のクリーク（ノードの集合において、集合内でノード全てが互いリンクし合っているような部分集合）のグラフの場合、二つ以上のクリークを合わせて一つのコミュニティとしたときに、モジュラリティが最大となることがある。その結果、クリークなど密な集団いくつかコミュニティ内にまとめられてしまう。

しかし、コミュニティとして抽出された個々のサブグラフの中だけに注目すれば、必ずサブグラフに対して、相対的に密な集団を抽出できると我々は考えている。そこで、再帰的にコミュニティ抽出操作を繰り返す手法を試みた。

具体的には、まず通常通りグラフ全体に対してコミュニティ抽出を行う。ここで得られた各コミュニティを新たなグラフとみなし、それぞれに対して再度コミュニティ抽出を行う。あるサブグラフに含まれるノードが一つになる場合と、どう分割しても Q 値が負になる場合に分割を停止するという方法をとった。その結果、図1に示すような階層的に分割されたコミュニティの樹形構造が得た。

3.3 可視化

階層的なクラスター構造を視覚的に把握するために、我々は、[10]で Hyperbolic, Radial and Space-filling (HRS) という可視化手法を提案した。HRS は定められた円の中でクラスターの階層構造を表示できる上、compound graph へ拡張を自然に行える。このため、解析結果の階層構造と元々のリンク構造を同時に示すことができる。本研究では、この可視化技術を使って、ソーシャルブックマークサービスのタグから得られるコミュニティの階層構造とその繋がりを表した。

表1 5つの最大コミュニティ内のトップ3タグ
Table 1 Top 3 tags for the five largest communities.

#1	technology	: fun, education, news
#2	toread	: business, work, productivity
#3	design	: art, blog, inspiration
#4	web	: software, tools, reference
#5	health	: food, recipes, cooking

表2 頻度が高いタグのトップ10
Table 2 Top 10 most frequent tags.

#1	web
#2	design
#3	reference
#4	tools
#5	software
#6	blog
#7	programming
#8	free
#9	imported
#10	cool

4. 実験

4.1 データ

本論文では、delicious のデータを対象とし、2008 年1月から4月にかけてクローリングを行った。得られたデータは、ユニーク URL 数が 38,783,521、ユーザ数が 1,469,051、ブックマーク数が 110,351,923、ユニークタグ数が 4,787,294 である。図2には、ユーザの登録数、URL の登録数、URL に付与したタグ数の分布を示したものである。これらの分布はベキ分布である。

この図から分かるように、多くのユーザが登録している URL の数が少ない (図2(a))。例えば、URL の 77% は一回しか登録されていない。同様に URL の多くは少数のユーザのみが登録している (図2(b))。例えば、ユーザの 39% は5つ以下のブックマークしか登録していない。多くのブックマークが少数のタグで登録されている (図2(c))。例えば、ブックマークの 75% は3つ以下のタグしか付与されない。この結果は、リソース中心のアプローチをとる時は注意が必要であることを示す。

得られたデータからタグの共起グラフを生成した。同時に登場している頻度を、タグ間のリンクの重みとして扱った。その結果、ノード (タグ) 数は 4,787,294、リンク数は 870,149,918 本となった。

また、これらのタグの中には一般的でない言葉 (自分の整理用に付けた番号や記号の羅列、複数の単語の並列、入力ミスなど) が膨大に含まれているため、Wiktionary^(注4)に記載されている約 20 万単語でのフィルタリングと、重み2以上でのフィルタリングを行った。そのうちの最大連結成分を取り出すことで、最終的に 106,182 ノード、33,743,680 リンクからなる重みつき

(注4) : <http://www.wiktionary.org/>

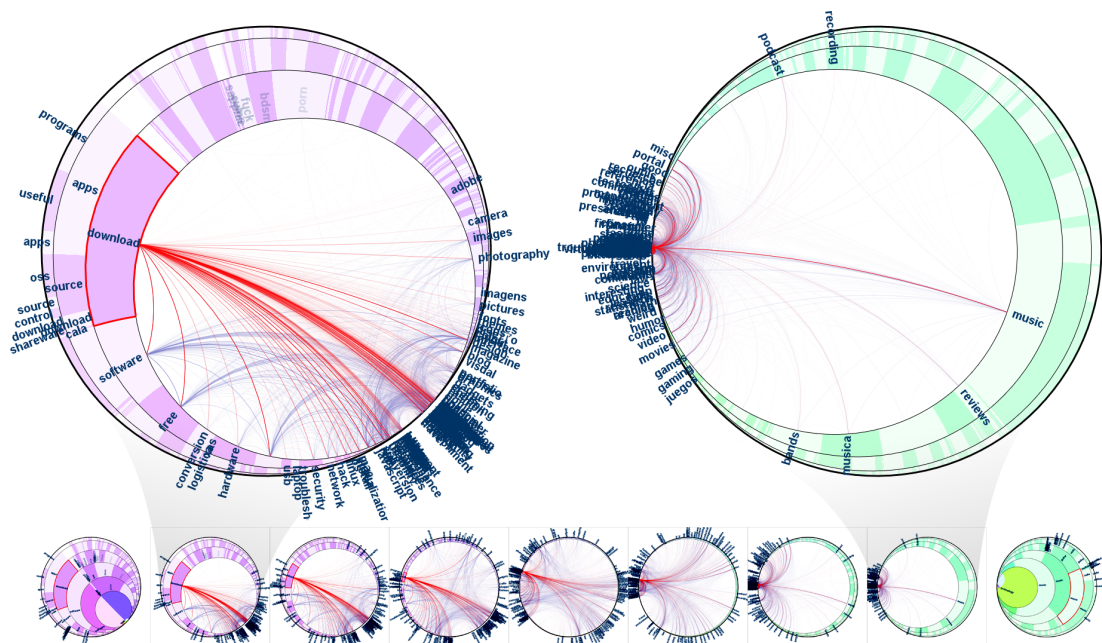


図4 可視化ツールにより，download と異なるコミュニティに所属する関係が強い単語の発見
 Fig. 4 Strongly related words to *download* from different communities was identify by the visualization tool.

文 献

- [1] N. Ali-Hasan and L. Adamic. Expressing social relationships on the blog through links and comments. In *Proc. of International Conference on Weblogs and Social Media*, 2007.
- [2] S. Bao, G.-R. Xue, X. Wu, Y. Yu, B. Fei, and Z. Su. Optimizing web search using social annotations. In *Proc. WWW '07*, pp. 501–510, 2007.
- [3] A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review E*, 70:066111, 2004.
- [4] L. Danon, A. Diaz-Guilera, and A. Arenas. The effect of size heterogeneity on community identification in complex networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2006(11), 2006.
- [5] S. Fortunato and M. Barthelemy. Resolution limit in community detection. *PNAS*, 104:36, 2007.
- [6] L. Guo, S. Jiang, L. Xiao, and X. Zhang. Fast and low-cost search schemes by exploiting localities in p2p networks. *Journal of Parallel and Distributed Computing*, 65(6):729–742, 2005.
- [7] P. Heymann, G. Koutrika, and H. Garcia-Molina. Can social bookmarking improve web search? In *WSDM '08*, pp 195–206, 2008.
- [8] Y. I. Leon-Suematsu and K. Yuta. A framework for fast community extraction of large-scale networks. In *Proc. WWW '08*, pp 1215–1216, 2008.
- [9] X. Li, L. Guo and Y. Zhao. Tag-based social interest discovery. In *Proc. WWW '08*, pp 675–684, 2008.
- [10] 村里英樹, レオン末松豊インティ. 双曲型多層円グラフ及び, 再帰的コミュニティ抽出法による, 大規模関係性データの視覚化手法. ネットワーク生態学シンポジウム, 2009.
- [11] M. E. J. Newman. Fast algorithm for detecting community structure in networks. *Physical Review E*, 69(6):066133, 2004.
- [12] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113, 2004.
- [13] S. Sen, J. Vig and J. Riedl. Tagommenders: Connecting Users to Items through Tags. In *Proc. WWW '09*, pp 1215–1216, 2009.
- [14] 湯田聡夫. 人の複雑ネットワークと情報ソムリエ: コミュニティ抽出法の可能性. 日本ロボット学会誌 26(1), pp.28–31, 2008.